Folksonomies to Support Coordination and Coordination of Folksonomies

Corey Jackson, Kevin Crowston, Carsten Østerlund, and Mahboobeh Harandi
School of Information Studies, Syracuse University, 343 Hinds Hall, Syracuse, New York, USA 13244 *(Email:{cjacks04,crowston,costerlu,mharandi}@syr.edu)*

Abstract. Members of highly-distributed groups in online production communities face challenges in achieving coordinated action. Existing CSCW research highlights the importance of shared language and artifacts when coordinating actions in such settings. To better understand how such shared language and artifacts are, not only a guide for, but also a result of collaborative work we examine the development of folksonomies (i.e., volunteer-generated classification schemes) to support coordinated action. Drawing on structuration theory, we conceptualize a folksonomy as an interpretive schema forming a structure of signification. Our study is set in the context of an online citizen-science project, Gravity Spy, in which volunteers label "glitches" (noise events recorded by a scientific instrument) to identify and name novel classes of glitches. Through a multi-method study combining virtual and trace ethnography, we analyze folksonomies and the work of labelling as mutually constitutive, giving folksonomies a dual role: an emergent folksonomy supports the volunteers in labelling images at the same time that the individual work of labelling images supports the development of a folksonomy. However, our analysis suggests that the lack of supporting norms and authoritative resources (structures of legitimation and domination) undermines the power of the folksonomy and so the ability of volunteers to coordinate their decisions about naming novel glitch classes. These results have implications for system design. If we hope to support the development of emergent folksonomies online production communities need to facilitate 1) tag gardening, a process of consolidating overlapping terms of artifacts; 2) demarcate a clear home for discourses around folksonomy disagreements; 3) highlight clearly when decisions have been reached; and 4) inform others about those decisions.

1 Introduction

Members of highly-distributed groups face particular challenges in achieving coordinated action. We examine how these challenges are addressed through the development of shared language and conceptual schema in the context of a web-based citizen science project. Citizen science is a broad term describing scientific projects that rely on contributions to research from members of the general public (i.e., citizens in the broadest sense of the term) who volunteer time and effort to advance the goals of the project. There are several kinds of citizen-science projects: some have volunteers collect data, while others, including the one we examine in this paper, have volunteers analyze already-collected data. The interactions

between volunteers and project organizers increasingly take place via the Web, e.g., on a site that accepts contributed data, or that presents data to be analyzed and collects volunteers' annotations (e.g., www.zooniverse.org). As a result, volunteers can be located anywhere in the world, with no particular organizational affiliation. Their work is sometimes described as "crowdsourcing science" and so is of interest and relevance to CSCW researchers.

The present paper focuses on citizen-science projects in which volunteers classify images, a common citizen-science data-analysis task. For example, in the *Snapshot Serengeti* project (www.snapshotserengeti.org), volunteers identify the species of animals shown in photographs taken by camera traps in the Serengeti National Park. The work, in this case, is the task of classifying an image: what species of animals are visible. In most citizen projects, such analysis work is entirely routine: the platform provides an image to a volunteer to classify, and the volunteer picks one or more labels to apply from a set that the project scientists include on the platform and perhaps answers some additional questions. The coordination required is the volunteers agreeing on the label or labels to be applied to each image. There is a simple pooled dependency among the volunteers' tasks so that the final answer can be determined simply as the consensus of the labels applied.

A few citizen-science projects are exploring less routinized citizen-science work, empowering volunteers to take part in the production of new knowledge rather than limiting them to pre-determined classification choices. However, the needed coordination is not well understood for this kind of work by these highly diverse and distributed groups. We are particularly interested in the role of shared language as volunteers work to develop new categories for classifying images. Such volunteer-developed categories form what are called folksonomies. Folksonomies play a dual role: supporting the volunteers as they work to label images at the same time that individual work leads to the development of a folksonomy. Accordingly, the research question we address in this paper is twofold:

First, how do folksonomies support coordination of non-routine classification work in online citizen science projects?

And conversely, what coordination is needed to create and maintain folksonomies?

2 Theory

Our paper draws on three bodies of prior research. First, we briefly review the CSCW literature on coordination to situate our discussion of folksonomies as a support for and result of coordinated work. Second, as we are interested in the mutual constitution of citizen-science work and folksonomies, we draw on structuration theory for the overall framing of our theorizing. Finally, after setting the stage with structuration theory, we discuss more recent research on folksonomies.

2.1 Coordination in CSCW

Coordination of interdependent work has been a perennial topic in CSCW research. Many CSCW systems have an explicit goal of supporting coordination. For example, an early CSCW system, the "Coordinator", sought to improve coordination by making communication more explicit about the coordination required (Flores, Graves, Hartfield, & Winograd, 1988; Winograd, 1987). Others have examined how particular system features support coordination. For example, Dourish and Bellotti (1992) argued for the importance of passive awareness to enable group members to work together. Dabbish et al. (2014) similarly noted that transparency, meaning system-enabled visibility of details of organizational processes or functions, is helpful for coordination. Many CSCW researchers have analyzed how coordination is achieved in various real-world settings. For example, Kittur & Kraut (2008; 2010) examined coordination in Wikipedia and wikis more generally.

Coordination is so central to CSCW that Schmidt & Simone (1996) describe coordination mechanism as a "conceptual foundation" for CSCW system design. Their analysis of coordination focused on how systems can support the articulation work needed to restrain complexly-interdependent activities. They defined a coordination mechanism as a coordination protocol for articulating interdependent activities, objectified in some artefact, e.g., a bug reporting form that reflects and shapes the interdependent activities in the procedure for tracking and fixing bugs.

Malone and Crowston (1994) similarly analyzed group action regarding actors performing interdependent tasks to achieve some goal, where the tasks might require or create various resources. In this view, actors face coordination problems arising from dependencies that constrain how tasks can be performed. In contrast to other theories that consider dependencies among actors, coordination theory classifies dependencies as occurring between a task and a resource, among multiple tasks and a resource, and among a task and multiple resources. In this perspective, the volunteers in an image-classification citizen-science project face a "shared-output dependency": their classifications are pooled to identify the consensus classification, meaning that the volunteers should agree on the appropriate label.

In developing the coordination theory framework, Malone and Crowston (1994) described coordination as relying on other necessary group functions, such as decision making, communications and especially the development of shared understandings and collective sense-making. In the citizen science case, volunteers need to share the set of appropriate labels and their meanings. Schmidt & Simone (1996) similarly note the locality of coordination artefacts within the particular social context in which they have meaning. However, neither work explores how such meanings are developed.

Other research in CSCW has touched on this question. For example, Ngwenyama & Lyytinen (1997) applied a social action framework to analyze groupware systems, identifying different categories of interaction. They noted that "communicative activity presupposes a common language, media, and a shared understanding of the organizational context" (p. 77) and that when communicative action fails, participants "shift either to discursive or strategic action to restore

understanding" (p. 77). Crowston & Kammerer (1998) analyzed software requirements groups and found that a well-developed collective mind, which includes shared representations, was helpful in supporting coordinated work. Menold (2009) identified the importance of cooperative work of shared technology frames, i.e., participants' common assumptions about information technology, and suggested ways to support the development of these frames. Yasuoka (2015) described the creation of project jargon as specialized professionals with their language collaborated. While these studies have demonstrated the importance of common language and some of the processes that build it, they include less discussion of how shared language supports coordination.

In summary, there is agreement that coordination is a key CSCW-system function and that coordination requires shared language and meaning. Similarly, some research has examined how such language is developed and its general importance for collaborative work. However, there is less research on the mutual constitution of shared language and coordination, the way shared language can be both a guide to and a result of coordinated work, our focus in this paper.

2.2 Structuration theory

To conceptualize the dynamic process by which individuals' actions relate to and influence others through shared language, we adopt a structurational perspective (Giddens 1984). We chose this framework because it provides a way to conceptualize how the behaviors of one citizen science volunteer might shape the actions of others, thus enabling their work to be coordinated. Structuration theory posits a recursive relation between team structure (defined as the rules and resources that influence, guide or justify individual action) and the actions of those that live within, and help to create and sustain, this structure. It is perhaps best described as a metatheory: that is, rather than specifically prescribing particular factors and actions or their relations, it describes the form that a theory might take. Specifically, structuration theory suggests that a theory of coordination in self-managing distributed groups should conceptualize structure and action in these group and describe how interrelation of these two achieves coordinated action.

In this paper, we consider structure as comprising three kinds of rules and resources identified in prior research (Barley & Tolbert, (Barley and Tolbert 1997; Stein and Vandenbosch 1996): (1) interpretive schema that create structures of signification, (2) authoritative and allocative resources that create structures of domination, and (3) norms and rules that create structures of legitimation. It should be noted that this division into three kinds of structure is an analytic convenience: in practice, they are overlapping and mutually reinforcing. Individual actions may be guided by these kinds of structure or may seek to change them. For example, an individual group member may follow a given process for a task (an individual action) because that process is the accepted norm within the group (i.e., because of a structure of legitimation).

Structure matters because the development of shared structure improves group performance if it enables more effective contributions by group members. It is not a question of the presence or absence of structure, but rather the nature of the

structures and the degree of agreement among group members. For example, without common interpretive schema (a kind of shared structure), individuals with different backgrounds may interpret tasks differently, making collaboration and communication difficult (Dougherty 1992). For example, a project like Snapshot Serengeti would not work if every volunteer had their own definition of animal species. And yet, the tendency for individuals to interpret tasks according to their perspectives is exacerbated when working in a highly-distributed environment, with its more varied individual settings and less opportunity for informal discussion and mutual observation.

We turn now to the question of how structure is developed. In some cases, structures may be set by external forces (e.g., the definition of species given by a zoologist to a citizen science volunteer). For this paper, we are interested in cases where they are emergent (e.g., the description of a novel species from the point of view of working zoologists). The key notion, in this case, is the "duality of structure", meaning that the structural properties of a social system are seen as both the means and the ends of the practices that constitute the social system. As Sarason (Sarason 1995) explains, in structuration theory:

> The central idea is that human actors or agents are both enabled and constrained by structures, yet these structures are the result of previous actions by agents. Structural properties of a social system consisting of the rules and resources that human agents use in their everyday interaction. These rules and resources mediate human action, while at the same time they are reaffirmed through being used by human actors or agents. (p. 48).

Simply put, by doing things, we create the way to do things (or as (Askehave and Swales 2001) put it more poetically, "the wheels of life go round, and as they go round, they form ruts which channel the wheels of life"). For example, the norm of using a particular process for a task is not a given, but rather is itself the outcome of prior actions by group members. By following the norm, members reinforce its legitimacy ("we always do it this way"); by taking different actions (e.g., skipping a step because it is seen to be too time-consuming or using a different approach because the accepted approach seems unable to deal with important problems), they undermine its legitimacy, perhaps eventually changing the norm.

Figure 1, adapted from Barley and Tolbert (Barley and Tolbert 1997), graphically summarizes the relation between institution (which the authors use synonymously with structure) and action, and how both evolve. In this figure, the two bold horizontal lines represent "the temporal extensions of Giddens' two realms of social structure: institutions and action," while the "vertical arrows represent institutional constraints on action" and the diagonal arrows, "maintenance or modification of the institution through action" (p. 100). In this figure, the influence of a norm on a member to use a particular work process is represented by a downwards vertical arrow, while reinforcement or changes to the norm due to actions is represented by an upwards diagonal arrow. We use this model of action and structure as the basis for our theorizing about coordination of work in highly-distributed groups.

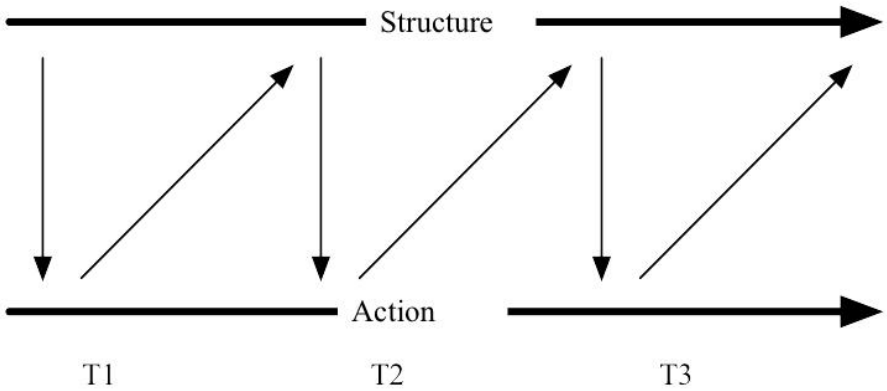


Figure 1. A sequential model of the relation between structure and action (from (Barley and Tolbert 1997).

In applying a structurational perspective to our analysis, we follow the lead of numerous authors who have similarly framed empirical analyses of team activities (Barley 1986; DeSanctis and Jackson 2015; Newman and Robey 1992; Orlikowski 1992; Walsham 1993). The perspective has been useful in particular for studying the development of virtual teams (e.g., (Sarker and Sahay 2003) and for CSCW research more generally (e.g., Kriplean et al., 2007; Nagar, 2012; McIntosh, 2008). For instance, Nagar (2012) examined the discussions of Wikipedia editors as they negotiate, interpret and construct the meaning of polices on the platform. The authors argue that while meaning and interpretations of policy are not always shared, members commit to "temporary and impartial interpretations" which become codified in policy pages. McIntosh (2008) similarly examined the recursive relationship between Wikipedia's neutral point of view policy (NPOV) and the production of news stories in Wikinews. The NPOV, an ethos document designed for Wikipedia articles, shapes new stories. However, editors are continually engaged in conflicts arguing the applicability of the NPOV policy, which in turn, has hampered the production of news articles.

While structuration has been applied to CSCW, these analyses have studied discussions around entire policies. There seems to have been less application of this perspective to the study of the development of folksonomies and accompanying interpretive schema, which is our focus in this paper.

2.3 Folksonomies

We are particularly interested in the role of shared or partially-shared language in helping groups achieve coordination. In the citizen-science context specifically, we are examining the role of an interpretive schema that guides citizen-science volunteers in applying labels to images. The dynamic relationship between structure and action can be seen especially in the use and evolution of social tagging functionalities common in social-software applications (e.g., social-bookmarking tools, social networking, photo and video sharing). One the one hand, highly-structured knowledge organization systems (KOS), e.g., thesauri or other classification systems applied by professional indexers offer a steady and nuanced

vocabulary and semantic structure. On the other hand, many sites support the rich and dynamic generation of freely-chosen user labels. Bringing the bottom-up creation of vocabulary closer to the formality of top-down classification schemes, we find folksonomies, a portmanteau of folk and taxonomy (Peters & Weller, 2008), referring to user-defined classification systems.

Folksonomies can emerge when users in a social environment label content and so need to reach agreement and consistency in the use of the labels. Figure 2 summarizes the core elements in this process as we conceptualize it. Starting in the upper left, by placing labels on individual items, participants seek to assign them meaning. As they label multiple items, they may seek to be consistent in the labelling so that the labels can serve to connect related objects. Through this process, individuals gradually create their own 'personomies', that is, a categorization system unique to their practices (upper right). Multiple purposes can drive the development of personomies, as individual labels do not necessarily refer to the content of the objects but can also denote author, origin, data form, work process, or other characteristics of the object salient to the individual.

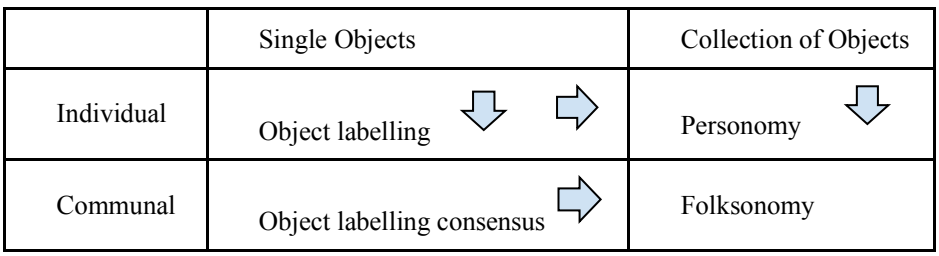


Figure 2. The evolution of folksonomies from individual labels.

At the collective level, multiple users examining the same object may strive to reach agreement about the most appropriate label to apply to describe that object (moving from the upper to the lower left quadrant in Figure 2). For example, a common use of labels is to enable content created by one user to be found by interested others, which requires agreement between producer and consumer on a label. This agreement might come through discussion about the particular object or more directly when one individual mirrors the observed practices of others, and so applies the same label. When visible to others the individual practice of applying labels may be seen as constituting a structure of signification that guides others.

Finally, folksonomies take shape from the compilation of personomies as consensus emerges around the appropriate set of labels and their meaning to guide collective labelling of the individual objects. In Figure 2, we show arrows from the other processes leading into folksonomies, but we note that the bottom arrow could as well run in the other direction: a developed folksonomy provides the basis for members of a group to achieve coordination in labelling, as it provides a structure of signification.

Of course, a free labelling practice may stall before achieving the shared vocabulary and semantic norms that constitute a folksonomy. Peters and Weller (Peters and Weller 2008) offer the metaphor of "gardens", where each label in a folksonomy is like a plant growing wild. A few labels may receive a lot of attention,

but often many others proliferate, yielding an unruly and overgrown garden in which it is hard to identify the important members. For instance, in a study of folksonomies, Al-Khalifa and Davis (2007) found that many labels overlapped by being 1) spelling variants or acronyms), 2) synonyms, 3) broader or narrower terms or 4) comparable thesaurus descriptors. Such a profusion of labels makes the collection less useful as a resource for achieving consensus on labelling objects.

To manage the messy nature of emerging folksonomies, Peters and Weller (Peters and Weller 2008) propose a number of "gardening" techniques that re-engineer folksonomies to make them more productive at the collective level. For example, many sites use word clouds to call attention to popular labels to help guide individuals in selecting one. Weeding is another strategy, in which misspelled labels or other closely-related labels are clustered together. However, these strategies by themselves do not solve the underlying problem in a folksonomy, the difficulty of making a coherent whole out of the many individual contributions.

Collective gardening is particularly challenging. Sometimes small groups can develop and maintain norms for labelling behaviors. However, larger communities often struggle to maintain such structures. Here, officially recognized folksonomy administrators could help implement effective gardening strategies. For instance, by combining folksonomies with more structured knowledge organization systems, administrators can add semantic structure to the folksonomy (Angeletou et al. 2007). As an example, Peters & Weller (Peters and Weller 2008) suggest the use of "power tags" as a starting point for gardening. Power tags are loosely defined as a small number of labels with particular importance to the folksonomy. Around those power tags, administrators can gradually perform their gardening. Another strategy is to separate personomies and folksonomies by requiring participants to keep their individual and communal-oriented labels separate. In the structurational framework, these strategies can be seen as efforts to identify particular approaches as authoritative and to create norms about the use of terms, that is, to create structures of domination and legitimation that reinforce the structure of signification.

In summary, the concepts of folksonomies, personomies, and gardening offer a framework to understand the labelling practices on citizen-science sites and similar crowdsourcing platforms. From this basis, we can see the tension between, on the one hand, the need for steady and complex knowledge-organization structures to describe complex data and, on the other, the freedom and flexibility of labels added dynamically by participants. A structurational perspective on folksonomies clarifies how they can be mutually constitutive of action (guiding labelling while also being formed by labelling) and also illuminates the need for other mutually-supportive structures for legitimation and domination so that others feel the need to use them.

3 Methods

In this section, we discuss the methods we adopted to answer the research question stated above, starting with a description of the research setting, then data elicitation and analysis approaches

3.1 Research Setting: Gravity Spy

The setting for this research is an online citizen science project called Gravity Spy (Zevin et al. 2016), hosted on the Zooniverse (Simpson, Page, and De Roure 2014) platform. The goal of Gravity Spy (www.gravityspy.org) is improving the instruments used to search for gravitational waves in the Laser Interferometer Gravitational-Wave Observatory (LIGO) scientific collaboration. The high sensitivity of the detectors need to detect evidence of gravitational waves means that they are also extremely sensitive to noise, both external (e.g., earthquakes or vehicular traffic) and internal (e.g., parts of the instrument interacting in unexpected ways). When such noise is recorded by the interferometers (called a "glitch"), it potentially blocks detection of gravitational wave signals, so understanding glitches and removing their sources is a key activity to improve the sensitivity of the detector. Having collections of glitches of the same class is useful to the LIGO engineers as they seek to locate and remove the cause of the glitches. To that end, volunteers perform two tasks, labelling glitches as members of existing glitch classes and identifying possible new glitch classes.

Classifying Glitches. The primary task volunteers perform in Gravity Spy is to classify glitches. The interface for classifying glitches is shown in Figure 2. Volunteers are provided with a glitch, represented as a spectrogram (on the left), a visual representation of the glitch that shows its intensity (represented by colour, from blue to yellow) at different frequencies (the y-axis) over time (the x-axis). The scientists and engineers who manage the interferometers have identified twenty-two classes of glitch, which are provided as options to the volunteers: these are shown to the right in Figure 2. Clicking on an option in the list brings up an image of a prototypical example of that class. Each class of glitch has a distinctive noise profile and appearance as a spectrogram. A volunteer labels a glitch by selecting the matching class from the list based on the similarity of the spectrogram to the exemplars. The spectrogram shown in Figure 2 represents a "very loud" glitch, as indicated by lots of yellow at all frequencies and the lengthy duration.

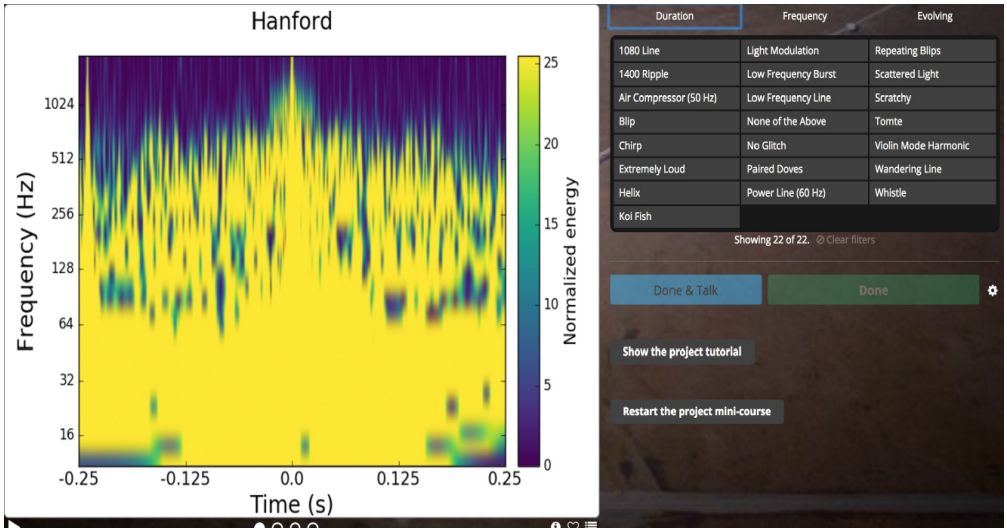


Figure 2. The Gravity Spy classification interface. Volunteers review the spectrogram on the left and select the glitch class on the right that fits best.

As of 2 March 2018, 11,427 volunteers (including members of the science team and the authors of this paper) have contributed 2,821,221 classifications of 439,265 glitches. Glitches are labelled by multiple volunteers and retired from the system when a consensus label is determined. At present, 105,574 glitches have been retired.

Identifying New Glitch Classes. Labelling glitches with the pre-defined glitch classes represent the lion's share of work in Gravity Spy. However, there are also glitches that do not fit a known class. It would be surprising to discover a new species of animal on the Serengeti (for example), but in contrast, the glitches in Gravity Spy are evolving as the LIGO detectors are worked on. Some issues are resolved, and those classes of glitch disappear from the data, but new kinds of glitches may emerge as the detectors change. Even in the current data, it is believed that there may be additional classes of glitches still waiting to be identified. Accordingly, in the primary labelling, when glitches do not fit one of the twenty-two known glitch classes, volunteers can label them as "None of the Above" (NoA).

To improve the system to handle these as-yet undescribed classes of glitches and so to better support the LIGO scientists, advanced Gravity Spy volunteers are invited to identify new classes of glitch. They do so by finding and describing sets of glitches with similar noise profiles and appearances that do not fit a pre-existing class. Volunteers can work independently or collaborate with other volunteers in the search for novel classes. The intent of describing novel glitch classes is that if a new noise profile is found to be common, the class can be added as a formal option in the main interface and, more importantly, the LIGO scientists can start to search for a solution. However, reaching agreement on exemplars, descriptions, and names for novel classes, that is, developing a folksonomy of novel glitches, is another shared-output coordination problem.

This kind of non-routine work is nearly unique among Zooniverse projects, so there is no explicit support for developing new glitch classes in the Gravity Spy infrastructure. Volunteers must instead re-appropriate existing system features to

support this process. One such feature is collections, which as the name implies, allow volunteers to collect objects of personal interest. After a volunteer classifies a glitch, it can be added to an existing or a new collection. Collections are named and can be private, public or shared with other volunteers. One way to document a possible new class of glitch is to develop a collection of examples.

In this paper, we focus on the coordination of non-routine volunteer work that takes place on Talk boards (i.e., discussion fora). There are multiple boards, created by the system developers to serve specific functions for the project. For example, on the Science board, volunteers can discuss the science of gravitational wave research and other research that might be pertinent to the interest of volunteers or scientists. The Help board is where volunteers ask questions about the interface or project. The Collections board is where volunteers discuss the collections they develop and search for collection collaborators.

Finally, and of specific relevance to our analysis in this paper, is the Note board. Discussion on this board is automatically linked to the individual glitches. Specifically, once a volunteer classifies a glitch, the system asks whether the volunteer wants to discuss the glitch with other volunteers. If a volunteer selects "Talk", they are taken to a thread in the Note board that includes any comments other volunteers have made about that specific glitch. Volunteers can post whatever they want: information related to how they classified the glitch, questions about their classification, or, of specific interest for this paper, hashtags for proposed labels to describe a new class of glitch. By hashtags, we mean a word (or several words run together) preceded by a # symbol that serves to label the particular object. Hashtags were not originally a feature of the Zooniverse Talk boards, but volunteers started using them after they became popular on other systems (especially Twitter) and support for hashtags has been improved (e.g., the system now lists popular hashtags and hashtags can be searched). We examined the use and evolution of hashtags for evidence of the development and use of personomies and folksonomies that reflect individual and shared language to describe glitches.

3.2 Data elicitation

This research employs the methods of virtual (Hine 2000) and trace ethnography (Geiger and Ribes 2011). Virtual ethnography is an approach that emphasizes researchers' participation in the online environment under study. Trace ethnography focuses on the archival or historical records (i.e., system logs) to construct the history of events as it appears in the system logs of the online environment. Data were collected as part of an ongoing research project designed to build citizen science projects that allow volunteers to conduct more complex analysis of data. Our data consists of notes from one year of participant observation, five semi-structured interviews, and system data from the project servers.

As virtual ethnographers, we created accounts on the platform and participated as regular volunteers classifying data, posting questions and responding to comments of other volunteers in the talk fora. At the time of writing, the first author reached Level 4, contributed 170 classifications, posted 10 comments, and created 3 image collections in Gravity Spy. Our role as researchers was known to

volunteers (the system shows a “researcher” label below our user ids on discussion posts), which could convey authority. Therefore, we were cautious about posting content that might influence debates about the accuracy of glitch classifications or name new glitch classes. Thus, while being aware of and taking note of these discussions, we largely remained passive, contributing only general comments such as “Yes. It does look like a reverse chirp”.

Our participation as volunteers allowed us to understand controversies and challenges in reconciling new glitch classes among many volunteers. As researchers, we observed conversation threads as volunteers debated hashtags, glitch similarities, proposed new glitch classes, and questioned the accuracy of their classifications with other volunteers. We also took stock of the unique and unexpected assemblages of technical features (e.g., collections and hashtags) that volunteers created to support their work. We discussed these activities in our weekly research meetings with scientists, researchers, and Zooniverse software developers.

As trace ethnographers, we examined the system log data to construct a history of interaction on comment threads. A significant advantage of studying trace data is that we can see in fine detail how language emerges and changes over time and how usage moves from individuals to shared. We focused our analysis a period ranging from April 2016 to February 2018. The trace data consist of the verbatim comments posted by volunteers and recorded by the system. Each comment included the unique id of the author and a timestamp. We used the trace data to visualize and better understand the provenance of terms or the evolution of specific hashtags that appeared in volunteer discussions. As an example, when a hashtag is proposed to name a candidate glitch class, using the trace data, we could document the history of the hashtag, the first appearance, the volunteer who suggested the hashtag, and subsequent comments using the hashtag. In that sense, trace ethnography, as a compliment to virtual ethnography, helps piece together events in the project.

Finally, we conducted two rounds of semi-structured interviews. We first conducted two interviews with scientists in two other citizen-science projects. The purpose of these interviews was to understand how projects with similar technical arrangements grapple with the use of hashtags in their projects. Second, we interviewed five expert volunteers in Gravity Spy. Expert volunteers were selected because they were heavily engaged in the classification task and were active contributors to social spaces in the project. Several of the expert volunteers were also engaged as project discussion moderators posing answers to questions, posting informative resources, and communicating with the science team. The goal of the interviews was to gain insights into current strategies volunteers enact to develop new glitch classes and to shed light on how the work of expert volunteers is currently supported by the current Gravity Spy infrastructure. Each interview lasted approximately one hour and was recorded and transcribed.

3.3 Data analysis

Using structuration as a theoretical lens, we analyzed interviews, participant observations and trace data. The data were discussed during weekly meetings in which we took stock of important themes around volunteer participation in the Gravity Spy project. Since our primary goal in the project is to develop infrastructure to support more complex citizen-science work, we discussed themes related to how volunteers currently experience the classification task and how they grappled with suggesting new glitch classes. In the findings, the trace data is used to support emergent themes in the results that emerged from volunteers building up evidence for new glitch classes.

In addition to this qualitative analysis, we conducted a quantitative analysis of hashtags focusing on their emergence and adoption by members of the project. Working from the traces of the Talk logs, we identified every use of a hashtag in a post, determining the frequency of use, the period over which the tag was used and the volunteers who posted the hashtags. In the findings, the data are primarily descriptive in that they provide illustrative examples of the trajectory of a particular hashtag. From this data, we can see who and during which periods volunteers used a particular hashtag.

4 Findings

4.1 Coordination in Classifying Glitches

As noted above, the primary task in the Gravity Spy project is classification of glitches into pre-defined glitch classes. The coordination problem in this task is ensuring that the distributed and heterogeneous group of volunteers arrive at the same classification for the glitch they observe. As noted above, the classification task is straightforward, and its work is routine, with a simple pooled dependency.

The science team who developed Gravity Spy provides tutorials and information resources to guide volunteers to classify glitches consistently. For example, the project website provides a field guide (Figure 3) that contains detailed descriptions of each glitch class: why it is important to the LIGO scientists, in which instruments the class of glitches appears, the visual characteristics of the glitch, and an image of a prototypical example of the class. With this information, volunteers can match the descriptions of classes in the field guide to the spectrogram they are classifying.

In the structurational framework, the formal classification system (the twenty-two pre-defined glitch classes) is a structure that guides the volunteers' classification work. It is at first and foremost an interpretive schema that creates structures of signification, identifying the kinds of glitches that are known to exist and that are relevant to improving the detectors. But in addition, the science team has made the decision about which glitch classes should be embedded in the site for the volunteers to classify. Thus, the tools also embody authoritative resources that create structures of domination, as the volunteers' activities are constrained limited to the options available in the interface. The instructions and interface

additionally create rules about how the classification should be done that create structures of legitimation, showing how the work should be done.

The classification task is controlled by the science team and cannot be directly modified by volunteers. The classification schema constitutes a formal knowledge-organization structure (KOS) leaving no room for adjustments by the volunteers at this level of engagement. As a result, we do not see a dynamic relationship between volunteers' actions and structure: rather, it is the work of the scientists and developers that have created the structures that guide the volunteers, and the volunteers can only choose to work within these structures or not contribute.

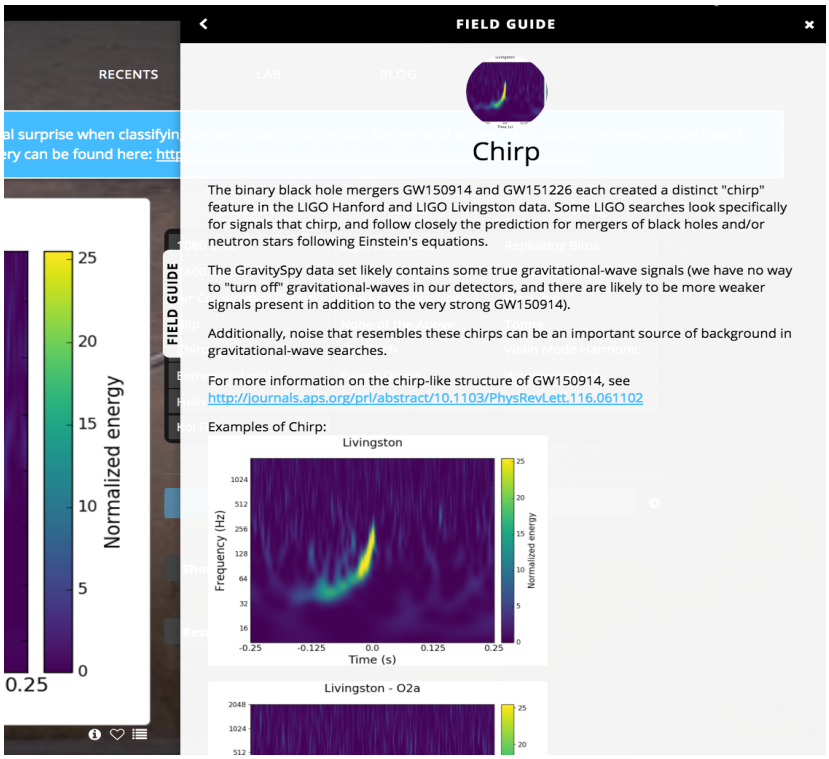


Figure 3. Available resources for the classification task. The field guide helps volunteers identify the glitch classes in the classification interface.

4.2 Structures for Coordinating Identification of New Glitch Classes

When a glitch does not fit one of the twenty-two glitch class options in the classification interface, volunteers should select "None of the above" in the interface. In additional, they may label the image in the Notes discussion board, either using an existing hashtag (i.e., drawing on a personomy or folksonomy to guide their work) or by creating a new hashtag that better describes the noise profile (i.e., possibly contributing to the evolution of a personomy or folksonomy). As other volunteers classify the image, they may also post a comment or add a hashtag to the image. However, the labeling performed in this step is problematic for several reasons. First, remembering the name and unique characteristics of each noise profile in the project can be challenging (currently there are 2,247 unique hashtags). Second, volunteers might unwittingly create new hashtags when one has already

been created by another volunteer(s) to describe the noise profile. Third, volunteers may create their own tagging systems that they might be intent to preserve (i.e., preferring a personomy to a folksonomy).

In this section, we focus on the varied coordination mechanisms enacted by volunteers that can develop folksonomies from these diverse idiosyncratic tags, viewing these as involved in the creation and maintenance of structures. Our analysis follows the outline in Figure 2 above. We start with individual object labelling that forms the foundation for personomies. At the communal level, individual hashtags may conflict and need to be resolved. We focus on the conversations taking place between and among volunteers as they build interpretive schema through their posts on images, enact authorities and allocative resources in building artifacts to support their hashtags, and creation of norms and rules through discussions about the homogeneity of different noise profiles. We see an informal process where volunteers create personomies and then engage in consensus building to structure a folksonomy.

Since coordination is a social activity, we first describe the social spaces in the project. In the system logs, we found 40,715 discussion threads comprising 68,969 comments, of which 71% (N = 29,390) are isolates, i.e., threads with only one comment. Tagging is an important activity in the project: 54% (N = 36,973) of posts contain at least one tag. As shown in Table 2, the most popular board is Notes. However, discussion takes place most often on other boards, e.g., Collections and Science, which include fewer volunteers and discussion threads, but more interaction, as indicated by the higher number of comments per thread.

Board Name	No. Users	No. Threads	No. Hashtags	Comments per Thread
Notes	1,390	40,211	36,563	1.62
Help	179	185	96	6.15
Chat	118	170	156	7.17
Science	58	75	65	9.72
Bug Reports	53	34	10	8.44
Collections	35	36	106	11.56
New Glitch Classes	8	4	4	8.5

Table 2. Basic statistics describing the discussion boards in Gravity Spy.

4.2.1 Personomies

Image Tagging. After classifying an image, volunteers may visit the Notes page belonging to the image and pose questions, leave general comments, or apply hashtags. Figure 4 shows an example of a post comment a glitch that includes general comments and hashtags. The general comments provide valuable insights into labelling practice, as volunteers often explain the logic of their decision to use a particular tag, e.g., by describing the noise profile and identifying characteristics of the glitch. For example, in Figure 4, the volunteer commented, "*Ive seen quite a few of these low frequency bursts that have vertical, linear holes"*. "Low frequency" describes the location of the noise and "vertical, linear holes" describes the form of the glitch. The volunteer also left a tag indicating that the glitch might

be an instance of a new class (#possiblenewglitch), and another suggesting a name for the glitch. When we examined the system logs, we found that *#snowmanburst* was applied to 48 glitches by this volunteer over the course of two weeks, but this hashtag was used only by this volunteer.

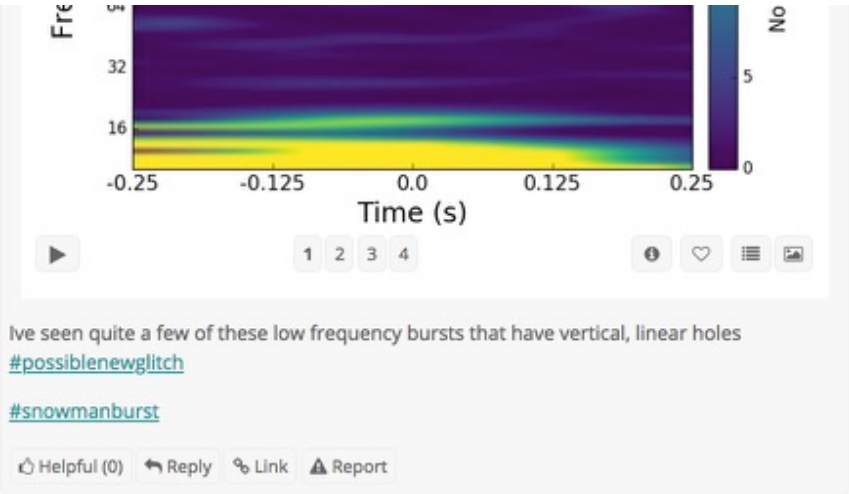


Figure 4. A comment posted by a volunteer describing the noise profile of a potential new glitch the volunteer named #snowmanburst. The volunteer describes the glitch, noting it exists at a low frequency and contains "vertical, linear holes".

The repeated, yet solitary use of a hashtag indicates the development of a personomy, as one volunteer chooses individually which hashtags to use for a glitch. As shown in Table 3, we found that 57% of hashtags were used by only a single volunteer. Nevertheless, we do observe volunteers developing classification schema to organize potential new glitches. In particular, some volunteers post detailed descriptions explaining their thinking, as shown in Figure 4.

No. of Volunteers	No. of Hashtags (% of total)	μ use in Days (sd.)
1	1,707 (57.3)	2.3 (9.49)
2–4	830 (27.9)	8.51 (20.13)
5–9	256 (8.5)	24.71 (37.76)
10 or more	148 (4.9)	114.76 (244.38)

Table 3. Descriptions of tag use by volunteers. Most hashtags are used by only one volunteer and are active in the project for only a short period.

Volunteers also apply hashtags as a higher-level organization schema to describe the overall characteristics of glitches. For example, labelling every image that has a noise profile that subsides over time with the #descending hashtag can help identify a commonality across glitches that may be the basis for identifying a new class. This practice is evident in a comment posted on a Notes thread, "Looks a bit like an #eiffel tower at around -0.125 sec. The loud #lowfrequencyburst at 0 sec is #descending." While not a glitch name, #descending helps link other noise profiles with similar features so volunteers can easily access glitches described by others as sharing this characteristic. We find #descending and other high-level descriptive hashtags (e.g., #increasing, #loud, #weak, #paired, and #off-center) are

also used by many volunteers. Other popular hashtags include the names of glitch classes from the classification interface. Finally, we find a few hashtags that have become a part of the collective knowledge of the group, suggesting the development of a folksonomy.

4.2.2 Coordination around Consensus Building

While personomies help individuals organize glitches into potentially new classes, they are also problematic to volunteers in several ways. First, as noted above, the rationale for volunteers' organization schemas are not widely accessible. In some cases, volunteers simply post a tag without additional comments that would help others understand the reasoning behind the label. Even when there is an explanation, it may be hidden in comments on the Notes board. Since a glitch has to be handled by only a few volunteers before retirement, it is likely that only a handful (if any) volunteers will come across a post on the Notes board.

Second, even when volunteers create resources describing their personomies, other users might prefer their own hashtags, causing conflicting representations and disagreements about which tag is appropriate. Such disagreements need to be worked out if the group is to achieve a consensus on the appropriate name for the new class. In short, no inherent stability exists in personomies.

To address these problems, discussions about hashtags and the characteristics of noise profiles that define their use is necessary. However, the boards were not explicitly designed to support this kind of coordination, so volunteers have taken to several workarounds to reach consensus. Through our observations of the conversations, we find two ways in which volunteers build consensus around glitches: *communicating practice* and *negotiating meaning*.

4.2.2.1 Communicating Practice

Articulating practice is an important activity when dealing with content on the Gravity Spy discussion pages. While many comments on the project contain only hashtags or what seems like personomy terms, there are a number of cases where comments contain detailed descriptions of practice, e.g., providing justifications for why the glitch deserves to be included as a new glitch class, the characteristics of the noise profile of the glitch, and the tag itself, all giving volunteers access to an individual volunteer's tagging practice. For example, in the following snippets, we can see volunteers making postulations about the form of the glitch and providing descriptions pointing to the morphological characteristics of the glitch:

> Maybe a few things going on here. Most prominent at t= 0.0 is what looks like a #aircompressor glitch. There's a good bit of the #70Hz line glitch and starting at around t= +0.125, maybe a #60HzPowerLine
> - Volunteer 1, 2017–07–25 02:54:05 (UTC)
>
> What is a #1400ripple?
> - Volunteer 2, 2016–11–06 00:52:43 (UTC)

> #1400ripple - short little ripple at 1400 hz with faded puppet line at 1024 hz
> - Volunteer 3, 2017–02–16 10:32:14 (UTC)

The post by Volunteer 1 adds specific detail describing the reasoning for identifying the glitch naming specific periodicities where the glitch was observed. The post points to potentially three overlapping glitches in one spectrogram, i.e., #aircompressor and #60HzPowerLine (two existing glitches in the classification interface) plus #70Hz, a new glitch. The post identifies the observed visual cues about the exact location of the glitch revealing #aircompressor is most visible at t= 0.0 (the x-axis in the spectrogram) and #70Hz at t = +0.125. The next two comments were posted on the same thread and are examples of volunteers learning about glitches from one another. When Volunteer 2 asks about #1400ripple, Volunteer 3 responds with details that can help Volunteer 2 (and others reading the post) learn how to identify the glitch. Comments like those posted by Volunteer 1 and Volunteer 3 are important because they represent descriptions of practice which serve as communicative devices for other volunteers seeking to learn about the tagging norms, in a sense these and similar comments are descriptions of personomies volunteers developed. Posed questions (19% of posts contain questions marks) and their responses are important activities for realizing a personomy and recruiting other volunteers to join in a search.

The technical features of the site also played a role in communicating practice. As volunteers construct personomies through other kinds of artifacts (e.g., individually-curated collections), linking to these resources became valuable in conversations as visual representations of their personomies or those of other volunteers. We found 321 posts with links to Gravity Spy collections (https://www.zooniverse.org/projects/zooniverse/gravity-spy/collections). Such posts are important since they provide a stock of exemplar spectrograms from which volunteers can learn what others consider examples of a potential new class. As an example:

> "#tightfireplace collection [here](https://www.zooniverse.org/projects/zooniverse/gravity-spy/collections/mjtbarrett/tight-fireplace) All from Livingston. Best seen in frame 4 for all examples. All have a widespread scratchy looking background extending to lower frequencies and similar shaped glitch from ~ 0 Hz-45 Hz lasting for ~ 1 sec.. Almost all are slightly offset from the centre. #possiblenewglitch ![Example Alt Text](https://panoptes-uploads.zooniverse.org/production/subject_location/f961b497-c71f-4d9f-bec1–4eba77bfe4ca.png)"
> - Volunteer 4, 2017–03–27 21:36:18

In this post, Volunteer 4 adds additional context to the hashtag #tightfireplace beyond the textual description of the glitch noise profile (i.e., "scratchy looking background extending to lower frequencies and similar shaped glitch from ~ 0 Hz-45 Hz lasting for ~ 1 sec."). Linking to a collection of example glitches helps readers better understand the intended characteristics of #tightfireplace. At the time of writing, the collection contained twelve glitches. The volunteer also adds another link that reveals what she believes to be a prototypical image. The descriptions left by Volunteer 4 extend beyond simply pointing to the noise profile and glitch characteristics and are accompanied by additional materials that volunteers can consult to learn about #tightfireplace.

4.2.2.2 Negotiating Meaning

Volunteers engage in several activities to resolve competing views of structure in tagging. There are several issues that volunteers need to work through as they consolidate potential new glitches and decide which terms to codify into the project's list of actively-used hashtags. This process can result in conflicts since volunteers are often unaware of the hashtags of other volunteers and their efforts to promote hashtags into folksonomies. However, we find that volunteers are adept at resolving such conflicts. They rely on several tactics that help achieve a shared understanding around hashtags: engaging in discussions around the appropriateness of certain terms to describe glitches, relying on the opinions of others when they are uncertain about the appropriate naming conventions, tag gardening to reconcile overlapping hashtags, and relying on external resources to support their postulations.

Discussing Glitch Morphologies and Hashtags. The Collections, Science, and Chat, and more recently the New Glitch Classes Boards are spaces where we find volunteers engaging in consensus building conversation. When compared to Notes, which has more than 40,000 threads, the compactness of these boards makes it easier to identify important threads. Further, the descriptive thread subject names, e.g., "Two Separate blips in a frame?" helps volunteers determine the focus of the thread and whether it is important for them to read. On these boards, we find longer discussions involving fewer discussants compared to the Notes board. The small number of discussants, however, is typical of similar online platforms where coordination occurs (e.g., Nagar, 2012).

The thread in Figure 5 is an excerpt of a conversation on the Chat board among five volunteers. The discussion, titled "#nettedBand – Looking for input" is centered on building understanding about the morphological features that are characteristic of the proposed nettedBand class of glitch. In the first post, a volunteer links to an image and describes the noise profile of the glitch in reference to a known glitch option and then asks other volunteers what they propose to call the glitch. The volunteer has been tagging similar noise profiles as #noneoftheabove but wants to achieve consensus on the most appropriate name for glitches with this noise profile. Several volunteers respond with their opinions, having seen similar types of noise in other images. As other volunteers joined the discussion, they propose their tags for this type of morphology. One volunteer (post not shown) stated, "I also find this type similar to those which are collected as MF horizontal scratchy which is not a great name …I think #lacyband is much better, as it also shows the possible relations between #lace and #lacy." Two new volunteers join the discussion (2nd and 3rd post in Figure 5), revealing that they applied #nettedBand and #lacyband to described the glitch and one volunteer shares his collection (in a private message) of #nettedBand images. The volunteers then come to some agreement that #nettedBand is the most appropriate tag. When other volunteers come across this thread they can see the logic for the term #nettedBand. From the last comment, readers can also infer additional structure in the tagging system, since she suggests lace, lacy, and lacyband hashtags can be considered sub-categories of #nettedBand.

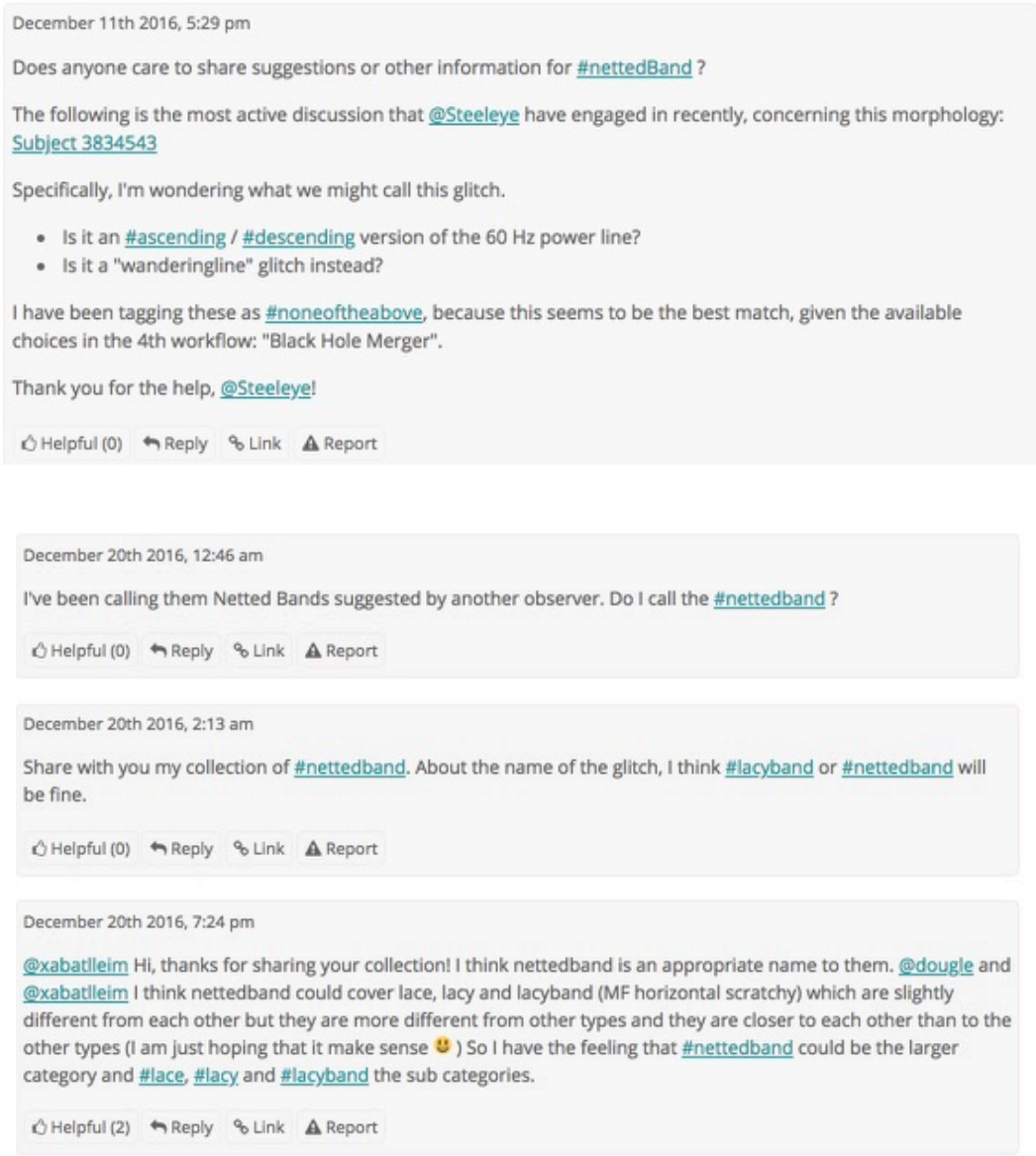


Figure 5. A discussion about the appropriateness of the name nettedBand for a class of glitch.

Relying on Collective Knowledge of Tagging System. Overt attempts at resolving glitches come in the form of a thread dedicated to a specific topic. Here more experienced members of the community serve as resources for newcomers and others. For example, one volunteer created a thread on the Collections board titled, "#lowfrequencysplatter: how should it be used?"[1]. This volunteer had created the #lowfrequencysplatter hashtag and wanted to engage other volunteers to determine what features of a glitch's noise profile define this glitch. In the thread, the volunteer shared several examples and concludes by stating, "I think it would be good if this discussion ends with either me expanding my definition to cover your

[1] https://www.zooniverse.org/projects/zooniverse/gravity-spy/talk/729/139098

use of it, you restricting your definition to meet mine, or both of us changing our definitions to meet in the middle for a common use of the tag".

Expert volunteers are important to advancing discussions because they possess knowledge of the many popular hashtags. The discussion below is an example.

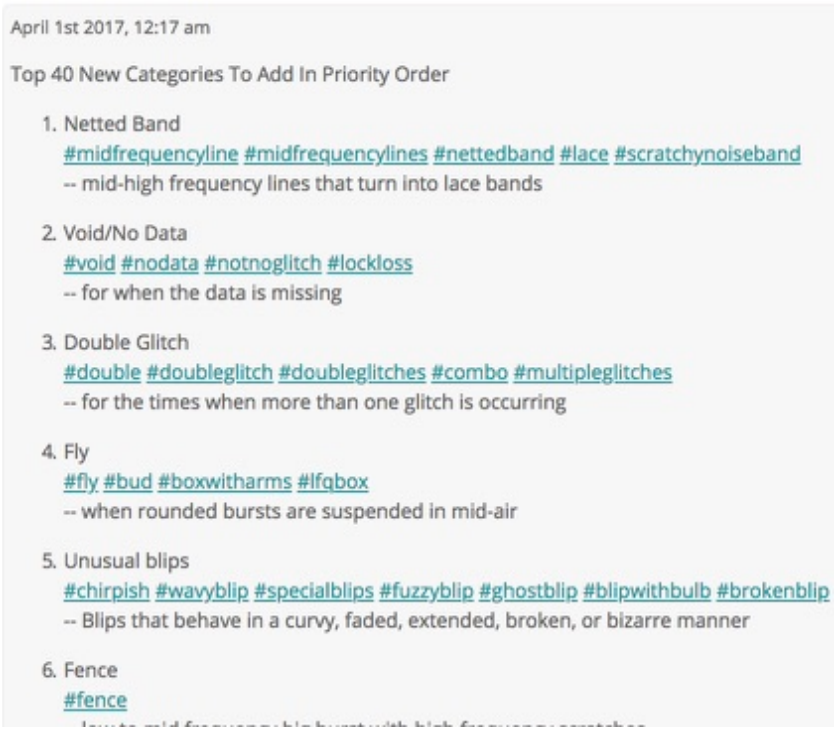
> @Volunter6 This is a #extremelyloud glitch framing in a 2 sec frame. Is not a 2 sec periodic glitch, but looks like a 2 sec framing case
> - Volunteer 5, 2017–09–28 09:16:19
>
> "@Volunteer5: the [exloud-koi-like collection] (https://www.zooniverse.org/projects/Zooniverse/gravity-spy/collections/melina-t/exloud-koi-like) by @Volunteer 7 contains many similar glitches to this example…"
> - Volunteer 6, 2017–09–28 12:31:54

In the thread Volunteer 5 (a newcomer to the for a, having been posting for only two months prior to this comment) calls on Volunteer 6 (a more experienced volunteer, having more than 1.5 years of fora activity at the time the comment was posted and a moderator at the time of writing) to question the characteristics of the glitch, offering pointers to the characteristics of the noise profile. Volunteer 6, having knowledge of active glitch searchers in the project, directs Volunteer 5 to the collection (and glitch *exloud-koi-like*) of another volunteer, Volunteer 7. While there is no formal hashtag being proposed, Volunteer 5 has learned other volunteers are searching for a similar profiled glitch.

Tag Gardening. The most obvious issue is competing names for a potential glitch. Tag gardening includes the process of consolidating hashtags. In Gravity Spy, we see volunteers taking stock of the existing hashtags, the noise profiles they represent, and their relationship to other hashtags. Frequently, volunteers propose to combine hashtags where they overlap or create sub-classes where variation among a set of glitches is relatively low. For example, several classes of glitch appear similar, but have slight variations in their noise profiles. In several discussion threads, volunteers worked to codify such known classes and sub-classes though posts on the fora that reference glitches, potential sub-classes, and link to collections. Peters (2009) points to hierarchical structures in tag gardening, where hashtags are synonyms. In the context of glitch identification, we find volunteers suggesting sub-classes of glitches, "I classified it as blip and label it as fly (I suggest that 'fly' of this type is a sub-class of blip)". The post by Volunteer 3 below is one example of volunteers building a "classification system" for the hashtags. Each list items contains the higher-level category, e.g., Netted Band, with associated hashtags, e.g., #midfrequencyline and a description of the category, i.e., mid-high frequency lines that turn into lace bands.

Another approach to gardening is combining overlapping vocabularies (see Figure 6). While volunteers individually understand the characteristics of the glitch to which they applied a hashtag, other volunteers might use a different vocabulary. To reconcile these vocabulary issues, volunteers alert others to similarities and overlapping vocabularies. For instance, in a Help board thread discussing similarities between #noiseband and other hashtags, one volunteer concluded "So #noiseband = #scratchy? A lot of the stuff under noiseband looks like #lace or something else though." Noting the inconsistencies in descriptions of noiseband,

another volunteer agreed that it should be discontinued, saying, "…in that sense noiseband is whatever you think it is. I personally would abandon it because it's not clear any more what they are and it isn't inherently obvious what the label really means."


April 1st 2017, 12:17 am

Top 40 New Categories To Add In Priority Order

1. Netted Band
   #midfrequencyline #midfrequencylines #nettedband #lace #scratchynoiseband
   -- mid-high frequency lines that turn into lace bands
2. Void/No Data
   #void #nodata #notnoglitch #lockloss
   -- for when the data is missing
3. Double Glitch
   #double #doubleglitch #doubleglitches #combo #multipleglitches
   -- for the times when more than one glitch is occurring
4. Fly
   #fly #bud #boxwitharms #lfqbox
   -- when rounded bursts are suspended in mid-air
5. Unusual blips
   #chirpish #wavyblip #specialblips #fuzzyblip #ghostblip #blipwithbulb #brokenblip
   -- Blips that behave in a curvy, faded, extended, broken, or bizarre manner
6. Fence
   #fence

Figure 6. Example of tag gardening where volunteer 3 suggests a hierarchy of glitch classes based on current glitches in the system.

Drawing on Resources. The materials created by volunteers in the project as well as LIGO-related resources, such as academic journals or alogs[2], are used as supporting materials in arguing for the existence of a new glitch class. 1,993 posts contain hyperlinks, of which 1,225 are to resources on the project website, e.g., links to threads on boards (e.g., 724 links to threads on the Notes board) or volunteer collections. However, links to external sites (e.g., scientific organizations or academic institutions) serve an important function offering authoritative sources of information as volunteers seek to figure out how glitches are created, the science behind the instrumentation, and interferometry. Our analysis revealed eighty-eight links to images on Imgur (a photo sharing community), thirty-two links to the LIGO domain (www.ligo.org), twenty-five entries pointing to alogs, and eighteen links pointed to sites in .edu domains. In one thread, a volunteer captured a screenshot of a spectrogram and annotated the figure to be used in a discussion about how spectrograms are generated pointing to concave zones and how the helical structure might correspond to a "sinegaussian envelop" altering how colors are assigned. In

[2] Alog is a digital notebook maintained by LIGO engineers that contain summaries of issues at the interferometer sites. A number of these issues are related to glitches. Available at https://alog.ligo-wa.caltech.edu/aLOG/

another thread, a volunteer posted a link to an alog entry posting, “Hey everyone, A note from the LIGO logbook about this. https://alog.ligo-wa.caltech.edu/aLOG /index.php ?callRep=23483”. The comment is part of a larger thread in which five volunteers discuss the power line glitch and what could potentially cause variations in its representation to support the creation of sub-glitch classes.

4.2.3 Maintaining Folksonomies

The activities described above help volunteers achieve shared understandings of glitch noise profiles and so come to some agreement on glitch naming conventions. While volunteers engaging in conversations on the boards achieve some understanding of the cadre of glitch classes that exists, a challenge remains in educating volunteers who do not participate on the boards. Additionally, the volunteers who have participated in discussions about glitch naming conventions need spaces to record resolutions, thus reifying a class’s position in the collective memory of the group. This resolution is achieved primarily through the development and maintenance of reifying artifacts that succinctly describe the glitches, their form, varied morphologies, prototypical examples. Attempts at recording these relations are spread across several posts but are consolidated by several expert volunteers in several threads particularly in “GravitySpy Classes – Collections”.

4.2.3.1 Creating Artifacts

Volunteers create artifacts on the site to reinforce norms around tagging, which make hashtags more resilient in the project. Not every hashtag is used to label a potential new glitch, so volunteers need to be aware of what hashtags are popular and what kinds of glitches are important to be discovered. Our observations pointed to several activities that volunteers perform to help popularize hashtags. Shown in Figure 7 (left) is a screenshot of a Collections post that groups hashtags based on a high-level abstract description, i.e., vertical repeating (twisted) types. Listing the hashtags and collections having prototypical subject images that could be associated with this abstract description could help others who are aware of the visual characteristics recall the appropriate hashtags to apply. The volunteer provides the scientists (and other volunteers) with detailed descriptions of how one might differentiate between the set of potential glitch classes, noting among other characteristics that a similarity between vortex and whirl is that “...both have max energy at 128 Hz.” Also important is the library of hashtags that currently exist in the project. Figure 7 (right) shows an excerpt of another post in Collections that lists all the relevant hashtags in the project. Clicking the hashtag in the list directs a volunteer to a search interface containing all images where volunteers applied the hashtag.

The structure of the post shown in Figure 7 (left) has been replicated across other posts and acts as a template for organizing hashtags. Acting as templates, these posts contain metadata elements used in the primary glitch classification task and additional elements that have contextual relevance to tagging glitches. These artifacts help reduce the need for other volunteers to develop their personomies.

The post shown in Figure 7 (right) is also important for the project since there is no single repository of hashtags maintained by the science team for volunteers to rely upon.

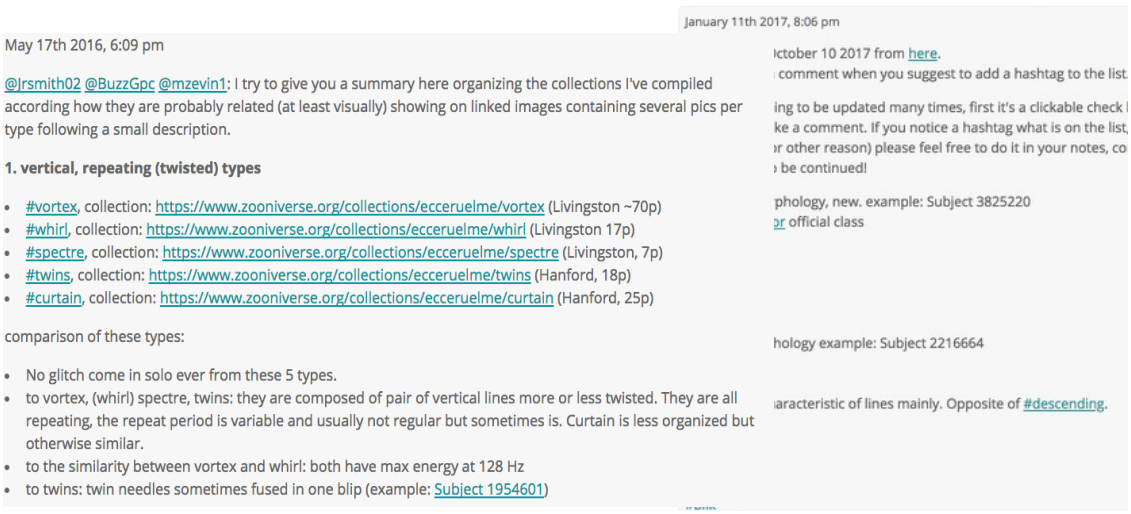


Figure 7. On the left is a screenshot of one volunteer's attempt to curate existing glitches and organize them by morphological features, e.g., vertical, repeating (twisted) types. On the right, a volunteer attempts to list the popular glitches in the project. Volunteers can see posts both posts on the fora.

5 Discussion

Online production communities and similar crowd settings create rich data that open a window into coordination available in few other settings. The paper contributes to the CSCW literature by exploring the dynamic relationship between coordination and emerging knowledge organization systems and vocabularies in crowdwork settings. Adding a structuration perspective allows us to approach individual labeling practices (i.e., personomies) and collective knowledge structures (i.e., folksonomies) as emerging out of a dialectic tension, which both facilitates the coordination of ongoing activities and require coordinative mechanisms to sustain a productive process. In our study, the trace data captured by the Zooniverse system enabled us to track how shared language, artifacts, and other key coordination support emerge as a result of collaborative work and conversely, how they guide and support such work.

Our data speak most clearly to our second research question, namely what coordination is needed to create and maintain folksonomies? In our data, we can see how volunteer actions support a developing structure of signification that enables volunteers to make sense of the ambiguous spectrograms they face on the system. Through engagement in conversations, volunteers develop shared understanding of glitches, their morphologies, the meaning of hashtags and when they should be applied. The shared language and artifacts emerging out of the participants' actions articulate collaborative structures in the form of interpretive schemas embodied in a hashtag folksonomy. However, our context exhibits an

additional step in the folksonomy-development process, namely personomies, individually-developed systems of hashtags. These uses of hashtags reflect an individual's point of view about particular glitches. But as personomies develop, volunteers sometimes create and share descriptions of potential new glitch classes, which over time can become shared.

As for our first research question, our data show that folksonomies support coordination of non-routine classification work in online citizen science projects by providing an emergent and evolving shared language with which to describe novel classes of glitches. The emerging folksonomy thus guides the on-going tagging. However, to be realized, volunteers must make visible and communicate these schemas to the broader community of volunteers. A problem is that the interpretive schemes are constantly evolving as volunteers encounter new data, making it difficult for others to keep up with tagging norms. It may not be clear to volunteers, particularly new volunteers, what parts of the folksonomy have stabilized and which are still evolving or indeed, are competing personomies. In other words, in this setting, we see overlap between and co-existence of collective and individual perspectives on the work.

From a structuration theory point of view, we should expect tension between productive and reproductive practice. Personomies give us a window into these evolving structures, as we find an ongoing tension between interpretive schemas generated through individual participant's practices and the collective interpretive schemas in the form of folksonomies. (Askehave and Swales 2001) note that "the wheels of life form ruts which channel the wheels of life". However, the ruts do not always form easily. It takes work to turn personomies into ruts that will not break the wheels of collective action.

Reflection on the study provides some more general take-aways for the three bodies of literature on which we based the study. The existing literature on coordination in highly-distributed groups often takes for granted the necessity of shared language and artefacts for coordination or focuses on the process of developing agreement and thus backgrounds how those agreements support coordination. Similarly, analysis of folksonomies often examines the creation of a folksonomy as a process separate from use. Our analysis of hashtags in Gravity Spy allows us to explore the dual nature of these coordinative processes. Folksonomies clearly emerge out of the citizen scientists' individual and collaborative practices. But, these emerging structures also guide their work.

The literature shows us that online communities often struggle to reach consensus and that structures, such as interpretive schemas, tend to emerge from small groups of high-status participants. In Gravity Spy, we also find a small group leading the development of a folksonomy. Of the approximately 11,000 volunteers who have contributed to the primary labelling task, only a small percentage participate in discussions at all (N = 1,448), and fewer than 200 take part in discussions beyond on the Notes board, where usage of hashtags can be synthesized.

However, the question of the status of the group is complicated by the citizen-science setting in which there is a clear and distinct status difference between the scientists who organize projects and the volunteers who work on them. Many

studies of folksonomies have been set in self-organizing groups, where those doing tag gardening can be given authority to do so (i.e., folksonomy administrators) or can take on such a leadership role for themselves. The volunteers contributing to the discussion of hashtags in Gravity Spy are active and viewed as leaders by others. However, it is difficult for any volunteer in a citizen science project to claim authority in the shadow of the science team. This question of authority is critical, as a key take-away from the application of a structurational perspective is how different structures are mutually supporting: it is difficult for interpretative schema to guide work if they are not supported by appropriate authoritative structures or norms.

5.1 Design Implications for Gravity Spy

The existing Zooniverse system on which Gravity Spy is built was explicitly designed to support the ongoing coordination of the work of labelling glitches as members of known classes. It does less well when it comes to coordinating the identification of new glitch classes and managing the productive tension between personomies and folksonomies. In this section, we discuss design implications that emerge from the research findings that suggest how to improve both the work of creating folksonomies and the role of folksonomies in coordinating work. As Lyytinen et al. (1992) point out, one important question for CSCW is how alterations to parameters of systems impact social interactions, below we list several recommendations based on this research.

First, consideration of the difficulties that volunteers faced in creating and maintaining folksonomies suggests the need for additional system features. Specifically, the current system lacks: 1) hashtag gardening tools and 2) identified places for discourse where common ground can be achieved. We also note the need to develop group practices that would support and be supported by these technical features.

Hashtag gardening tools. A lot of the work volunteers put into labelling unknown glitches end up wasted or serving personomies alone, as illustrated by the fact that 57% of hashtags are only used by one user. The existing system lacks tools for hashtag gardening. It is possible to search for glitches with a particular hashtag, but it is not easy to see what other hashtags have been used for those images and how one hashtag is related to others. There is thus no easy way to compare and contrast multiple personomies, e.g., to eliminate overlapping hashtags or to build hierarchical structures with sub-classes. Such tools have been implemented in other collective systems, though they are often restricted to a few power users.

Places for discourse. As noted by Ngwenyama & Lyytinen (1997) when communicative actions fail participants need to shift to discursive action to develop common language and understanding. However, communication in asynchronous settings can exacerbate this issue since users might not be able to update and adjust their perspectives (Rader, 2010). We see that volunteers do engage in extensive conversations about which hashtags are most appropriate for a potential new glitch class and suggest gardening to combine overlapping hashtags. However, it is not clear in Gravity Spy where the best place is for such discourse. Discussions about

appropriate hashtags are thus scattered across the system: in notes for different glitches or in separate discussions on other boards. The result is that discussion is fragmented and invisible to the majority of volunteers.

Group decision-making practices. A further problem in the project is that agreement does not necessarily emerge out of discourse among participants, as there is no set process for deciding on hashtags or even recognition that a folksonomy is a desired outcome of the project. Depending on their notification settings, volunteers may not even be notified if someone else comments on a thread after them. As a result, posts about a glitch are often fewer conversations than sets of sequential observations (or single observations: 71% of comment threads have only one post) and discussions are often inconclusive.

We also note limitations of the system for supporting coordination of non-routine classification work. The system lacks features to 1) help volunteers choose appropriate hashtags and 2) to use hashtags that have been applied to objects. Again, we also note the need to develop group practices that would support and be supported by these technical features.

Choosing hashtags. As noted above, discussions about hashtags for possible new glitch classes are spread across multiple boards and discussions. As a result, even when there is a consensus among a group of volunteers about an appropriate hashtag for a new glitch class, there is no easy way of making other volunteers aware of it. The website does display the twenty most-common hashtags, but most of these are the label of known glitch classes rather than new glitches. Nor is there any tailoring of the list to the particulars of the glitch to help guide a volunteer to an appropriate choice. We do see volunteers writing posts describing the characteristics of potential new glitch class to other volunteers, e.g., a detailed discussion thread titled "The Zooniverse Hashtag System[3]" in which volunteers convened to grapple with some of the issues noted above. However, the product of these conversations is only known to those who participate in or viewed that thread. The system lacks any way for volunteers to mark particular hashtags as being synonymous, preferred or deprecated or a way to record a definition of the meaning of a particular hashtag that can be easily accessed by all volunteers, functions that are found in other systems.

Pooling of labels. Automatic detection of vocabulary similarities has been proposed in other communities (e.g., Chen, 1994) to address term variation in asynchronous collaboration. However, unlike in routine classification work, Gravity Spy does not offer system feature that computes agreement among volunteers on labelling with hashtags, nor is there any push from science team for the volunteers to reach consensus. With the current free-for-all approach to labelling, it would be quite challenging for a system to tell when consensus had been reached. As a result, the hashtag labelling process may support the development of a folksonomy, but it does not support the LIGO scientists in improving the detector. Indeed, hashtags are not currently used by the science team. We suspect constructing a hierarchical taxonomy, e.g., Heyman & Garcia-Molina (2006) used principles from social network analysis to do so, might help show how

[3]https://www.zooniverse.org/projects/zooniverse/gravityspy/talk/329/121461?comment=216032&page=1

labels are related, providing scientists with glitch references that could help isolate the cause of certain glitches.

Enforced or expected use of hashtags. Structuration theory suggests that the development of a shared language and artifacts is interdependent with the emergence of norms and rules that create structures of legitimation and authoritative resources. However, we have noted a lack of authoritative resources and norms and rules, that is, structures of domination and of legitimation, which support the development of structures of signification. These kinds of structure are prominent in the main classification interface but absent from the advanced work of identifying new classes. As a result, volunteers feel no obligation to use—or even an expectation that they use—hashtags that others have developed. In short, in Gravity Spy, the link between creating and applying a folksonomy is tenuous: there is no drive or support for a discussion about the label for a particular object to converge and no straightforward way for a discussion about appropriate hashtags to influence labelling practice.

As part of our study of the nature of advanced work on Gravity Spy, we interviewed the organizers of other Zooniverse citizen science project that had volunteers engaged in advanced work. The work in Chimp & See (https://www.chimpandsee.org) presents an interesting contrast to Gravity Spy. In this project, volunteers tag videos of chimpanzees for behaviours, other species and the identity of the individual chimps in the videos. The project scientists collaborated with volunteers (e.g., in Skype meetings) to build a corpus of hashtags for the volunteers to use. The science team posts a list of the important hashtags they have curated and tutorials for applying the hashtags. The hashtag guide also points to several instances where hashtags are particularly valuable, e.g., "you've identified the animal, but the ID was especially hard, and/or you had to look at the neighboring videos in order to figure it out" and "you can't ID the animal (then add #need_ID)." In other words, this project has created authoritative resources for what hashtags to use and norms and rules about their use that reinforce the structuring power of the interpretive schema. Having the science team as an authoritative source managing and promoting folksonomies thus helps keep the use of hashtags under control and productive for the science, but at the same time perpetuates the status difference between scientists and volunteers.

6 Conclusion

From the results above, promoting more complex work in crowdsourced citizen science requires new infrastructure to support coordination of both creating and using a folksonomy. However, in the absence of such infrastructure, volunteers can appropriate existing system features to support their work. The results presented above reveal some ways that volunteers deal with the lack of technical infrastructure and guidance by more authoritative members of citizen science projects. To advance citizen science beyond current state-of-the-art in image tagging, we suggest new infrastructure to support more collaborative work.

First, online citizen science projects are designed primarily to support classifying existing data into pre-defined categories of interest to the science teams who run the projects. We believe that there is a potential for systems to increase in complexity to support more advanced work. For example, we envision voting mechanisms for individual subject images when there is disagreement about what hashtag to apply.

Second, we suggest the need for additional discussion spaces for developing structures of domination or legitimation to guide the application of the folksonomies. While the current discussion spaces are used by active volunteers, they are not well curated nor are they visible to the majority of volunteers. The case of Chimp & See shows how the science team can provide such structures. Yet we are interested in how advanced volunteers can also be empowered.

Accordingly, our final recommendation is that citizen science project managers should consider delegating more power to volunteers to make decisions as they re-envision the role of citizen scientists as scientific assistants. The empirical examples above point to the competence of citizen scientists to handle complex coordination roles in defining glitches and resolving conflicts. Additionally, when left to their own devices citizen scientists in other projects have been responsible for scientific discoveries (Lintott et al. 2009; Straub 2016). In our case, as volunteers produce new glitch classes, they help gravitational-wave physicists better understand the nature of their data. How to provide volunteers with the authority and legitimacy they need for this work is an open question, but one that must be answered for them to be able to make their non-routine work more productive.

7 Acknowledgments

We thank the many citizen scientists whose engagement have made Gravity Spy possible and our collaborators on the project, including M. Zevin, S. Coughlin, S. Bahaadini, E. Besler, N. Rohani, S. Allen, M. Cabero, A. Katsaggelos, S. Larson, T. Littenberg, A. Lundgren, J. Smith and L. Trouille. Gravity Spy is partly supported by the US National Science Foundation award INSPIRE 15–47880.

8 References

Al-Khalifa, Hend S., and Hugh C. Davis (2007). Towards better understanding of folksonomic patterns. *In Proceedings of the Conference on Hypertext and Hypermedia,* Manchester, UK, 10–12 September, pp. 163–166.

Angeletou, Sofia, Marta Sabou, Lucia Specia, and Enrico Motta. (2007). Bridging the gap between folksonomies and the semantic web: An experience report. *In Proceedings of the European Semantic Web Conference,* Innsbruck, Austria, 3–7 Jun, pp. 30–43.

Askehave, Inger, and John M. Swales (2001). Genre identification and communicative purpose: A problem and a possible solution. *Applied Linguistics*, vol. 22, no. 2, pp. 195–212.

Barley, Stephen R. (1986). Technology as an occasion for structuring: Evidence from observations of CT scanners and the social order of radiology departments. *Administrative Science Quarterly*, vol. 31, no. 1, pp. 78–108.

Barley, Stephen R., and Pamela S. Tolbert. (1997). Institutionalization and structuration: Studying the links between action and institution. *Organization Studies*, vol. 18, no. 1, pp. 93–117.

Black, Laura W., Howard T. Welser, Dan Cosley, and Jocelyn M. DeGroot. (2011). Self-governance through group discussion in Wikipedia. *Small Group Research,* vol. 42, no. 5, pp. 595–634

Chen, Hsinchun. (1994). Collaborative systems: Solving the vocabulary problem. *Computer,* vol. 27, no. 5, pp. 58–66.

Crowston, Kevin, and Ericka Eve Kammerer. (1998). Coordination and collective mind in software requirements development. *IBM Systems Journal*, vol. 37, no. 2, pp. 227–245.

Dabbish, Laura, Colleen Stuart, Jason Tsay, and Jim Herbsleb. (2014). Transparency and coordination in peer production. *arXiv preprint 1407.0377.*

Dougherty, Deborah. (1992). Interpretive barriers to successful product innovation in large firms. *Organization Science*, vol. 3, no. 2, pp. 1–25.

Dourish, Paul, and Victoria Bellotti. (1992). Awareness and coordination in shared workspaces. *In CSCW' 92. Proceedings of the 1992 ACM Conference on Computer Supported Cooperative Work,* Toronto, Ontario, Canada, 1–4 November, pp. 107–114.

Flores, Fernando, Michael Graves, Brad Hartfield, and Terry Winograd. (1988). Computer systems and the design of organizational interaction. *ACM Transactions on Office Information Systems (TOIS)*, vol. 6, no. 2, pp. 153–172.

Geiger, R. Stuart, and David Ribes. (2011). Trace ethnography: Following coordination through documentary practices. *In Proceedings of the Hawaii International Conference on System Sciences*, Kauai, HI, USA, 4–7 January.

Giddens, Anthony. (1984). *The Constitution of Society*. John Wiley & Sons.

Hine, Christine. (2000). *Virtual Ethnography*. SAGE Publications Ltd.

Heymann, Paul, and Hector Garcia-Molina (2006). *Collaborative creation of communal hierarchical taxonomies in social tagging systems*. InfoLab Technical Report 2006–10, Stanford. Available from: http://ilpubs.stanford.edu:8090/775/1/2006–10.pdf

Karsten, Helena. (2003). Constructing interdependencies with collaborative information technology. *Computer Supported Cooperative Work (CSCW),* vol. 12, no. 4, pp. 437–64.

Kittur, Ankit., and Robert Kraut. (2010). Beyond Wikipedia: Coordination and conflict in online production groups. *In Proceedings of the Conference on Computer Supported Cooperative Work,* Savannah, GA, USA, 6–10 February, pp. 215–224.

Kittur, Ankit., and Robert Kraut. (2008). Harnessing the wisdom of crowds in Wikipedia: Quality through coordination. *In Proceedings of Conference on Computer Supported Cooperative Work (CSCW),* San Diego, CA, USA, 8–12 November, pp. 37–46.

Kriplean, Travis, Ivan Beschastnikh, David W. McDonald, and Scott A. Golder. (2007). Community, consensus, coercion, control: Cs*W or how policy mediates mass participation. *In Proceedings of the Conference on Supporting Group Work,* Sanibel Island, FL, USA, 4–7 November, pp. 167–176.

Lintott, Chris J., Kevin Schawinski, William Keel, Hanny Van Arkel, Nicola Bennert, Edward Edmondson, Daniel Thomas et al. (2009). Galaxy Zoo: "Hanny's Voorwerp," a quasar light echo? *Monthly Notices of the Royal Astronomical Society*, vol. 399, no. 1, pp. 129–140.

Lyytinen, Kalle J., and Ojelanki K. Ngwenyama. (1992). What does computer support for cooperative work mean? A structurational analysis of computer supported cooperative work. *Accounting, Management and Information Technologies,* vol. 2, no. 1, pp. 19–37.

Malone, Thomas W., and Kevin Crowston. (1994). The interdisciplinary study of coordination. *ACM Computing Surveys (CSUR)*, vol. 26, no. 1, pp. 87–119.

Menold, Natalja. (2008). How to use information technology for cooperative work: Development of shared technological frames. *Computer Supported Cooperative Work (CSCW)*, vol. 18, no. 1, pp. 47

McIntosh, Shawn. (2008). Collaboration, consensus, and conflict. *Journalism Practice,* vol. 2, no. 2, pp. 197–211.

Nagar, Yiftach. (2012). What do you think? The structuring of an online community as a collective-sensemaking process. *In Proceedings of the Conference on Computer Supported Cooperative Work*, Seattle, WA, USA, 11–15 February, pp. 393–402.

Ngwenyama, Ojelanki K., and Kalle J. Lyytinen. (1997). Groupware environments as action constitutive resources: A social action framework for analyzing groupware technologies. *Computer Supported Cooperative Work (CSCW)*, vol. 6, no. 1, pp. 71–93.

Orlikowski, Wanda. J. (1992). The duality of technology: Rethinking the concept of technology in organizations. *Organization Science*, vol. 3, no. 3, pp. 398–427.

Peters, Isabella, and Katrin Weller. (2008). Tag gardening for folksonomy enrichment and maintenance. *Webology*, vol. 5, no. 3, pp. 1–18.

Rader, Emilee. (2010). The effect of audience design on labeling, organizing, and finding shared files. *In Proceedings of the Conference on Human Factors in Computing Systems*, Atlanta, GA, USA, 10–15 April, pp. 777–786.

Sarason, Yolanda. (1995). A model of organizational transformation: The incorporation of organizational identity into a structuration theory framework. In *Academy of Management Proceedings*, vol. 1995, no. 1, pp. 47–51.

Sarker, Suprateek, and Sundeep Sahay. (2003). Understanding virtual team development: An interpretive study. *Journal of the Association for Information Systems*, vol. 4, no. 1.

Schmidt, Kjeld, and Carla Simonee. (1996). Coordination mechanisms: Towards a conceptual foundation of CSCW systems design. *Computer Supported Cooperative Work (CSCW),* vol. 5, no. 2–3, pp. 155–200.

Simpson, Robert, Kevin R. Page, and David De Roure. (2014). Zooniverse: Observing the world's largest citizen science platform. *In Proceedings of the 23rd International Conference on World Wide Web*, Seoul, Korea, 7–11 April, pp. 1049–1054.

Straub, Miranda. (2016). Giving citizen scientists a chance: A study of volunteer-led scientific discovery. *Citizen Science: Theory and Practice*, vol. 1, no.1, pp. 2–10.

Walsham, Geoff. (1993). *Interpreting information systems in organizations*. John Wiley & Sons, Inc.

Winograd, Terry. (1987). A language/action perspective on the design of cooperative work. *In Proceedings of the Conference on Computer-supported Cooperative Work*, Austin, TX, USA, 3–5 December, pp. 203–220.

Yasuoka, Mika. (2015). Collaboration across professional boundaries: The emergence of interpretation drift and the collective creation of project jargon. *Computer Supported Cooperative Work (CSCW),* vol. 24, no. 4, pp. 253–276.

Zevin, Michael, Scott Coughlin, Sara Bahaadini, Emre Besler, Neda Rohani, Sarah Allen, Miriam Cabero, Kevin Crowston, Aggelos Katsaggelos, Shane Larson, Tae Kyoung Lee, Chris Lintott, Tyson Littenberg, Andrew Lundgren, Carsten Østerlund, Joshua Smith, Laura Trouille, and Vicky Kalogera. (2017). Gravity Spy: Integrating Advanced LIGO detector characterization, machine learning, and citizen science. *Classical and Quantum Gravity*, vol. 34, no. 6, pp. 064003.